Center for Information Technology

Leveraging Big Data to Advance Biomedical Research

Tuesday, June 18, 2019

For Federal News Network's digital transformation month in April, Andrea Norris, the Director of CIT and Chief Information Officer of NIH, sat down with Tom Temin on the Federal Drive with Tom Temin podcast. They discussed how recent enhancements to NIH's computational infrastructure, including supercomputing and big data, affect NIH's biomedical research.

Recognizing an incoming tsunami of data, NIH worked to modernize its network as a "broad enterprise infrastructure," according to Norris—meaning that the organization shares data across one network within their own facilities as well as through cloud computing systems.

In response to a need for massive computing power by intramural researchers to analyze data, NIH has invested in expanding Biowulf, one of the world's most powerful supercomputers. Since NIH's network modernization began, Biowulf's network traffic has doubled each year, making it one of the largest distributed 100-gigabit networks for research in the world. As a testament to its raw computing power and cutting-edge capabilities, Biowulf is ranked among the top 100 most powerful supercomputers in the world by the TOP500 project.

Biowulf is available exclusively to NIH's intramural researchers, and the only supercomputer designed and supported solely for biomedical research in the world, according to Andrea. As a result, NIH could not model the design of Biowulf after any other federal agency's system. "We provide a lot of very personalized support of expert folks who can help these researchers tune their algorithms, structure their data in a way that will optimize both the compute and the power of what they're doing analytically," Andrea said during the podcast.

Looking ahead to the next five years, NIH will further leverage the power of supercomputing and big data programs to enhance its mission: turning discovery into health. In the interview, Norris noted how research data is often compartmentalized and separated among researchers. This makes it difficult for other researchers to access that data, determine its usefulness in their own work, and then make it available for the next researcher. The need for efficient connectivity led to NIH's next era of scientific data research: cloud computing.

The most significant of NIH's cloud computing endeavors is The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. The STRIDES Initiative, a medium used to access cloud service providers (CSPs), was developed by NIH to "move the national biomedical research ecosystem forward in a cohesive way," Norris said. NIH has partnered with commercial cloud providers like Google Cloud and Amazon Web Services to reduce barriers to accessing and computing on large biomedical data sets.

“Cloud is going to be a very important part of [our] strategy in data science,” Norris said, referring to NIH’s Strategic Plan for Data Science. “It moves our national biomedical research ecosystem forward in a cohesive way to really leverage the power of compute, the power of the big data programs, and ability to generate and harness intelligence out of data that we could not do before today.”

Additional cloud computing projects NIH’s researchers are undertaking include the All of Us Research Program, the NIH HEAL Initiative, and the TOPMed program.

While there is optimism around the work already conducted using cloud computing, there are still areas to address. One of the biggest concerns for some of these projects is data security. For programs that collect identified data, NIH has made a tremendous investment into security and privacy protocols, as well as in ensuring personal data is only used in places where the participants give consent.

“[Most data] comes into NIH deidentified; we don’t even have the identification associated with it. That’s important for us,” Andrea said.